Statistical vs Analytical Significance: Is My Model Good Enough?

Introduction

Models and simulations (M&S) are becoming increasingly important throughout the acquisition life cycle of defense-related aerospace systems, both to justify selection of a single design from among competing alternatives, and to reduce the costs of system development and testing of a selected design. The credibility of the M&S used to perform these functions is, and will increasingly be, a major concern of defense planners. But the question of credibility raises even more concern about the cost of establishing that credibility: how much will I have to pay to find out if a model meets my needs? If the cost of verification, validation and accreditation (VV&A) is more than my project can afford, then can I get by with less? How much V&V is enough? How will I know? And can I convince someone else (like OSD) that I don't need more V&V ?

Partly to try and answer those questions, the Office of the Secretary of Defense (OSD) commissioned a five year project, the Susceptibility Model Assessment and Range Test (SMART), to develop a credibility assessment process for DoD and apply it to a set of important M&S. In the process of applying this process, the SMART project learned some valuable lessons about how to make M&S VV&A cost-effective, and how to decide how much V&V is enough.

The way to make VV&A cost-effective is two-fold: first, concentrate not on verification and validation activities, but on an assessment of how the model is going to be used to help solve your problem. In order to determine how much V&V is enough, first figure out what you're going to use the model to do, and how well you need the model to do it. Then you can compare what you already know about the model with those requirements, to see if (1) you already know enough about how well the model works that you don't need to do any more V&V, and (2) if you need to do V&V, which parts of the model are most critical to your problem so that you can focus your V&V effort on what's really important to you. By focusing on model accreditation requirements, a user can avoid spending resources conducting V&V that isn't really necessary for his particular application. Second, to make VV&A cost effective, report the results in a way that facilitates the determination of M&S suitability and make these results readily available so that the next user of that model doesn't have to start from scratch like you did. V&V results are most useful to the accreditation proponent when the implied strengths and weaknesses of the model are explicitly stated and the impacts on model usage are described. You can achieve both of these objectives by documenting your results in a standard format and making them available to the community through an existing information repository.

Modeling And Simulation And The System Acquisition Process

Within the DoD, models and simulations are used in support of system acquisition. From development of a mission needs statement, through concept development, demonstration, engineering & manufacturing, to production and deployment and operational support, M&S are being used today to not only define requirements for the system, but to design it and train military personnel in how to use it. These M&S are used to support Cost and Operational Effectiveness Analyses (COEA's), which support buy/don't buy decisions; they are used to support test and evaluation by explaining test results, extending test conditions, and predicting test outcomes; they are used by system developers to evaluate design alternatives; and they are used by the training community to provide synthetic environments in which to fight the wars that might someday arise.

Report Documentation Page

Report Date 00AUG2001	**Report Type** N/A	**Dates Covered (from... to)** -

Title and Subtitle Statistical vs Analytical Significance: Is My Model Good Enough?	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Naval Air Warfare Center, Weapons Division (Code 418000D)1 Administration Circle, China Lake, CA 93555	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es)	**Sponsor/Monitor’s Acronym(s)**
	Sponsor/Monitor’s Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes	
Abstract see report	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** SAR
Number of Pages 8	

Serving to underline the need for M&S credibility in the acquisition process, the Under Secretary of Defense signed out a new policy[1] on December 4, 1995 called “Cost as Independent Variable” (CAIV). This policy, which applies to the acquisition of new systems, and the upgrade of older systems, promises to have a significant impact on the way the Department does business. CAIV is a departmental acquisition strategy which entails setting aggressive, realistic cost objectives for acquiring defense systems, and managing risks to obtain those objectives. Cost objectives are set early in a program, balancing mission needs with projected out-year resources. Once the system performance and cost objectives are decided (on the basis of cost-performance tradeoffs using M&S), cost becomes more of a constraint, and less of a variable.

Until recently, DoD’s design processes have been largely driven by an unrelenting threat and by available technology, not always emphasizing cost-performance tradeoffs in setting program goals. The CAIV approach formalizes the process for cost-performance tradeoffs and better connects the user, supporter and developer, arriving at an affordable balance between cost, performance and schedule. Relying heavily on these tradeoffs, of course, puts an even greater burden of credibility on the M&S used in support of acquisition programs.

Aggressive cost control implies the existence of cost objectives that are the DoD-equivalent of sound commercial business practices. These objectives will be set as early as possible (e.g., Milestone I or before for most systems). It is expected that these cost objectives will be much lower than would be projected for a system using past ways of doing business. The reason for this emphasis on cost is apparent from Figure 1.

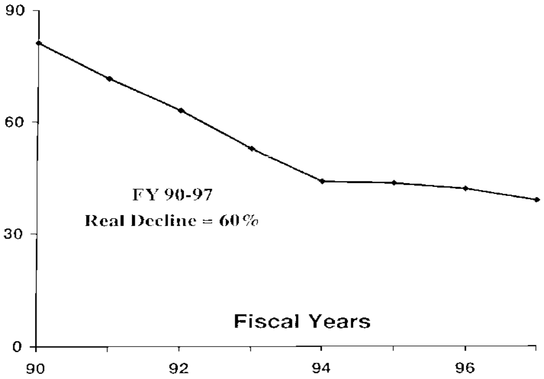

Figure 1. DOD Procurement Budget (FY97 $B)[2]

The reality is that we have less to spend on equipment to support our troops. If the needs are still there, but the money isn’t, then some equipment will not be bought. The DOD is saying that we have to cut our costs, if we are going to do our mission. Thus the total cost of a system, given a constant or declining budget, will be a strong factor in whether or not we buy it. For example, suppose we need a widget which is highly effective for a given task and we can specify a set of technical characteristics that this widget should have, but those technical characteristics result in a very expensive widget (this is a not uncommon occurrence in the DoD in the last several years). With a limited budget, our alternatives seem to be either to not buy it, or to cancel another program in order to get it. However, if a less expensive thingy with a

[1] USD(A&T) Memorandum, entitled “Reducing Life Cycle Costs for New and Fielded Systems” dated December 4, 1995

[2] Source: P-1 Budget Document

different set of technical characteristics can do the job (perhaps in another way), then we may have a third alternative. But the only way we'll find that third alternative is if we have some faith in the M&S that support the cost-performance tradeoffs that sort out the solution.

And therein lies the problem. We have only limited faith in many of the M&S that support the system acquisition process, mostly due to less than adequate resources applied to the VV&A of those M&S. The reasons for the hesitancy of acquisition programs to sign up for M&S VV&A are fear of cost and the basic uncertainty of what is really needed: how much V&V is enough, and how little V&V cost can I get away with?

Some Definitions

In order to answer those questions, we first need to be clear on some definitions. The basic elements of M&S credibility have been defined through the efforts of the Military Operations Research Society (MORS), and codified in JCS PUB 1[3]. They are:

> Verification: The process of determining that a model implementation accurately represents the developer's conceptual description and specifications.
>
> Validation: The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.
>
> Configuration Management: The process of applying technical and administrative oversight and control over the model.
>
> Accreditation: An official determination that a model is acceptable for a specific purpose.

In other words, verification means finding out if the model does what you think it does, validation means finding out how well it does it, configuration management means finding out if you've got the version you think you do, and accreditation means deciding that it's good enough for you!

It should be noted that under these definitions, validation is not an absolute end in itself; rather, it is a process of gradually shedding more and more light on the ability of the model to represent reality. The only absolute under these definitions is accreditation: that enough information is available to make the decision that this model fits a particular application's requirements.

Validation is the element which generally receives the most interest and scruitiny in the VV&A process, but it is also the most costly and elusive due to the difficulty and expense of obtaining adequate test data.

[3] Department of Defense. *Department of Defense Dictionary of Military and Associated Terms*, Washington D.C., DoD, 23 March 1994 (Joint Publication 1-02)

As a result, in many cases subject matter expert reviews, or qualitative assessments of the model's "acceptance in the community" have been the sole basis for accreditation decisions. Many acquisition programs suffer from "validation avoidance." The root of the problem is that validation is not always understood to be a process, rather than an absolute end in itself. The question: "Is your model validated?" leads inevitably either to never-ending validation, since there is no definition of what's good enough correlation with test data to decide you can quit doing it, or it ends with the user deciding that validation is too hard (or too expensive) so he avoids it altogether. Validation really only has meaning in the context of an application: the user wants to know if the model is demonstrated to be good enough for his purpose. This means that the user has to analyze his application, determine what "good enough" means for him, and only do enough validation to determine if the model meets his needs. In other words, the user must search for "analytical significance" in validation results.

Analytical Vs. Statistical Significance

A typical approach to detailed validation involves decomposing a model into functional elements. This allows for (a) identification of "testable" elements of the code, and (b) examination of how each of those elements contributes to the overall credibility of the model for "end-to-end" model comparison with test data. That is, if we compare the model's final output parameters with test data, and the results do not correlate well, we must have functional level results to sort out which part of the model is causing the problem (assuming that the problem is not in the test data). And, if the results do correlate well, we want to be sure that it was not due to a serendipitous cancellation of compensating errors within the model.

The breakdown of the model into "testable" functional elements supports the definition of validation data requirements; that is, in order to validate a model, we must translate model validation needs into test plans and test plan supplements to obtain the required data. Suitable data may be obtained from ongoing testing, but it requires pre-test coordination for data instrumentation requirements that may be unique to model validation issues. Data obtained from testing in this manner is at a low enough level in the model to support statistical comparisons between model predictions and test results. Often this is not the case when considering the model as a whole, since many model outputs are not directly measureable at a test facility, or may only be inferred from the data that are collected.

Figure 1 illustrates sample results from a test, as a function of time, when compared with model predictions under the same conditions as the test. This represents the comparison of a single model functional element, such as radar tracking performance, with actual test data. Because of the relatively small variation in the test data compared to the difference between test and model prediction, the difference clearly is statistically significant. Therefore, from a statistical standpoint, we should reject the hypothesis that the test data and the model predictions represent the same distributions. But what does that mean? That is, does the statistically significant difference between the test data and the model predictions really make any difference to the user? Are the test and model result differences "Analytically Significant?"

The answer to that question depends on how the model predictions are to be used in the application of the model, and what questions the user is trying to answer. What are the user's "critical analytical issues (CAI)" ? This is really the meaning of analytical significance: is the difference between test data and model prediction significant in my context? Would that difference change my final answer?

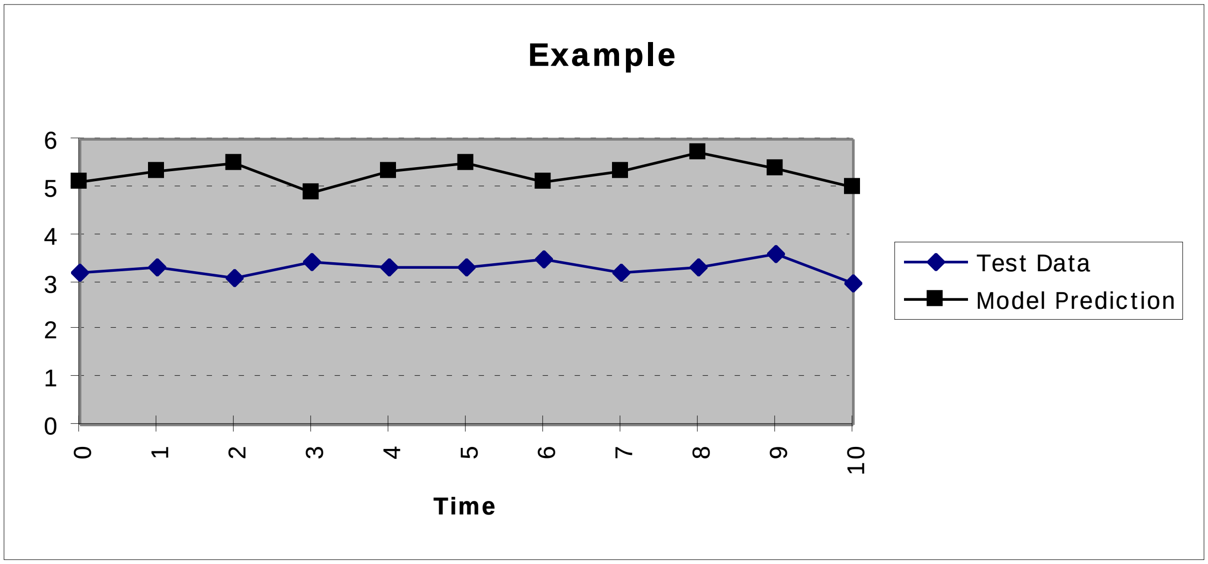

Figure 1. Sample Model Functional Element Correlation with Test Data

In many cases, model or test results are used to feed higher level models which address user critical analytical issues; in other cases "testable" model functional elements are directly embedded into higher level models whose results address CAI's. Figures 2 and 3 represent the results of processing the functional element data represented in figure 1 through two different higher level models. In both cases, the model functional element predictions were used for one curve, and the actual test data were fed into the model for the other curve. Also shown on figures 2 and 3 are acceptance criteria for the applications using the models. In figure 2, for the model addressing CAI #1, the difference between the model functional element results and the test data results are "analytically significant," because the differences in model output using those data fall outside the acceptance criteria boundaries. In figure 3, however, use of the same data through another model yields results that are not significantly different, even though their differences were statistically significant. In that case, we would say that this functional element of the model is sufficiently accurate to support the requirements of CAI #2, or that the model supporting CAI #2 is not very sensitive to the functional element we tested.

Another more concrete example of this idea is shown in figure 4. That figure illustrates the miss-distance distribution of a surface-to-air missile against an aircraft target, as measured by two different measurement systems, for two different test series. The set of data labeled "MTS" clearly exhibits a Rayleigh distribution, with a fairly well defined "cutoff" limit at the extreme. The data labeled "GPS", on the other hand, presents a Poisson distribution, with outlying miss distances well beyone the cutoff of the MTS data. A statistical analysis of these data concludes that they do not come from the same distribution, even though they purport to represent the same missile system against the same target. (This may be due to a number of factors, including differences in the way that the two trials were set up, as well as differences in the measurement systems themselves).

But what does that mean? The fact that these two miss distance distributions are statistically different does not, by itself, mean anything other than that they are different. The question that needs to be answered is "does this difference mean anything to my use of the data?" The data in question were intended to be used to support estimates of probability of kill (PK) of the aircraft target by the missile system. Using the distribution of miss generated by the "MTS" system resulted in a computed PK of 0.74, with a confidence interval of plus or minus 0.02; the PK computed using the "GPS" system was 0.77, with a confidence interval of plus or minus 0.03. Considering that the program using these data would have been happy with PK estimates accurate to one decimal place, the difference between these PK's is insignificant. Thus, even though the two distributions are statistically different, for this particular application they are equivalent in an analytical significance sense.

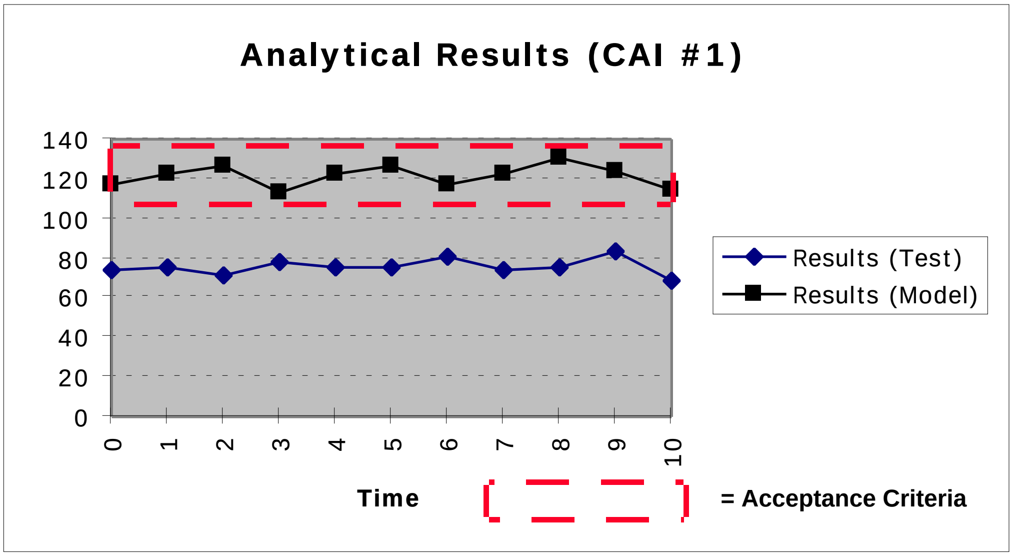

Figure 2. Model Level Results for CAI #1

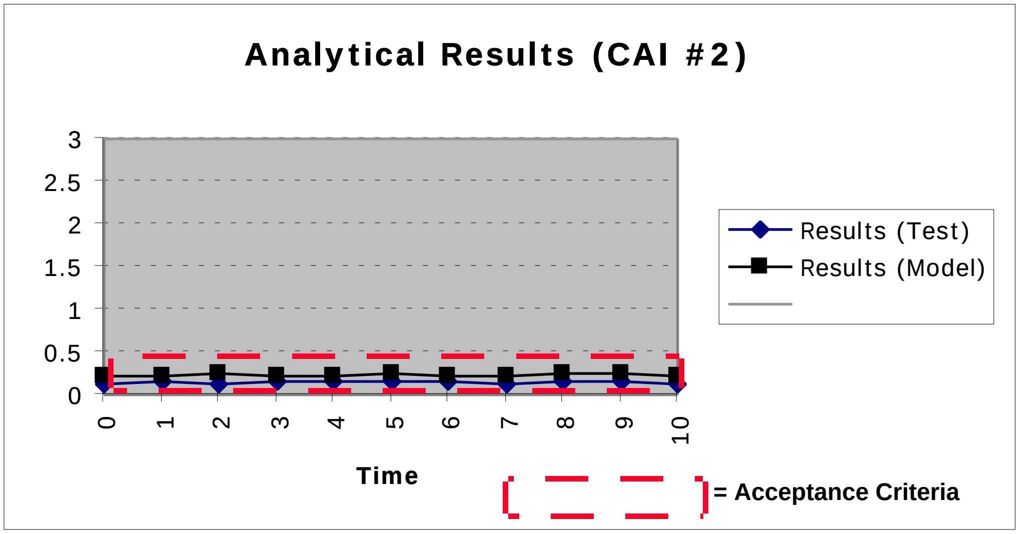

Figure 3. Model Level Results for CAI #2

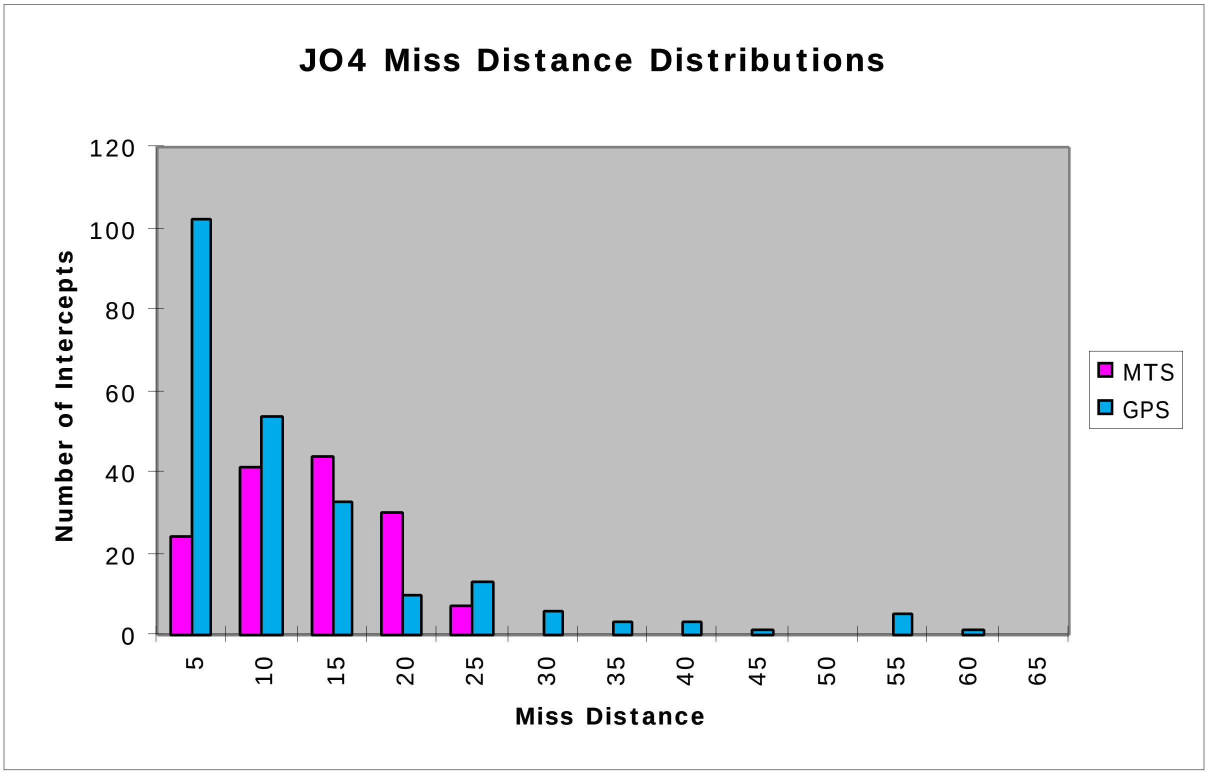

Figure 4. Miss Distance Distributions Resulting from Two Data Sources

Acceptance Criteria

But how do we determine the acceptance criteria boundaries shown in figures 2 and 3? The acceptance criterion in either case must come from fidelity requirements that are determined by an analysis that defines the "range of analytical significance" of the application. The key to making any V&V effort cost-effective is to tie the requirements for V&V activities to accreditation requirements. And those accreditation requirements come from the analysis of the application: What are the key study questions to be answered? What is the hierarchy of measures of effectiveness (MOE) that address those questions? How do those MOE's relate to model outputs? How do changes to those model outputs, and consequently to those MOE's, change the answers to my study questions? From the answers to those questions, particularly the latter, come M&S acceptance criteria for the particular application.

Acceptance criteria come in three flavors: functional requirements, fidelity requirements, and operating requirements. Functional requirements identify those elements of the problem that must be included in the model in order to satisfy the issues and requirements of the application. Fidelity requirements are driven by the sensitivity of the study questions to variations in the MOE's and model outputs. Operational requirements are driven by the computer hardware, manpower and funding resources available. It is these criteria which in turn drive the requirements for V&V activities in support of M&S accreditation.

Once those criteria are established, the model's capabilities can be compared with those criteria to determine its suitability for a particular application. The results of these comparisons are threefold:

(a) Data To Support M&S Accreditation (how the model fared compared to the acceptance criteria);

(b) A Risk Assessment (what areas of the model pose risks to the study questions?); and

(c) Requirements for Model Enhancements (what areas of the model should be "fixed" because of their high risk to the program?).

These are the results of comparing the model with test data for "analytical significance": it ties the entire process to the real requirements for the model, in the context of an actual application. Tests of statistical significance are really suited more to an assessment of the behavior of the test data, rather than their significance to the problem. We can use statistical techniques to evaluate the suitability of the data for use in comparison with model predictions, but statistics themselves will never answer questions about whether the model is suitable for a particular application.

Summary

The credibility of M&S used in the system acquisition process is crucial to informed use of those models in support of system design, T&E and training. The key to affordable VV&A activities is to focus V&V efforts on the part of the model that's important to your problem. Analytical significance is the key criterion for M&S comparison with test data, where the comparison is based on acceptance criteria that are determined by analyzing your M&S application. By taking this approach we can go beyond tests of statistical significance to tests of whether the model is good enough for what the user needs it to do. V&V for its own sake is never cost-effective, but V&V conducted to support a specific problem focuses those activities on key analytical issues and makes the process efficient and cost-effective. So ultimately, the answer to the question, "How much V&V do I need to determine if my model is good enough?" is "It depends on what you want the model to do!"